LINKS and MS2LINKS
USER MANUAL
v.03

Eizadora Yu, Andrew Rothfuss, and Malin Young
Sandia National Laboratories
Livermore, CA
https://ms3d.ca.sandia.gov
Updated: June 13, 2007

Table of Contents

I. Introduction

Overview

Structural elucidation of biomolecules and their assemblies is crucial in understanding the molecular basis of biological function. While high-resolution techniques like NMR and x-ray crystallography remain the premiere techniques, some biomolecular structures are still unknown, intractable to these techniques primarily because of challenges with size limitations and the dependence on crystal formation. To overcome these limitations, a growing number of alternative strategies that rely on sparse distance constraints (hydrogen/deuterium exchange (HDX),[1, 2] spin labeling,[3] fluorescence resonance energy transfer (FRET)[4, 5] have been employed to obtain structural information on such biomolecules.

In addition, mass spectrometric 3D (MS3D) approaches that couple chemical crosslinking and footprinting techniques with MS analysis has proven to be a valuable technique for the investigation of the 3D structure of proteins, nucleic acids, and macromolecular complexes.[6-16] The wide range and availability of crosslinking and footprinting reagents in conjuction with the flexibility of the MS analytical platform provides an excellent alternative technology for the structural elucidation of biomolecules that are not readily amenable to the traditional structural techniques. In this direction, new computational tools for the interpretation of mass spectra from crosslinked and modified proteins have been created.[9, 17-22]

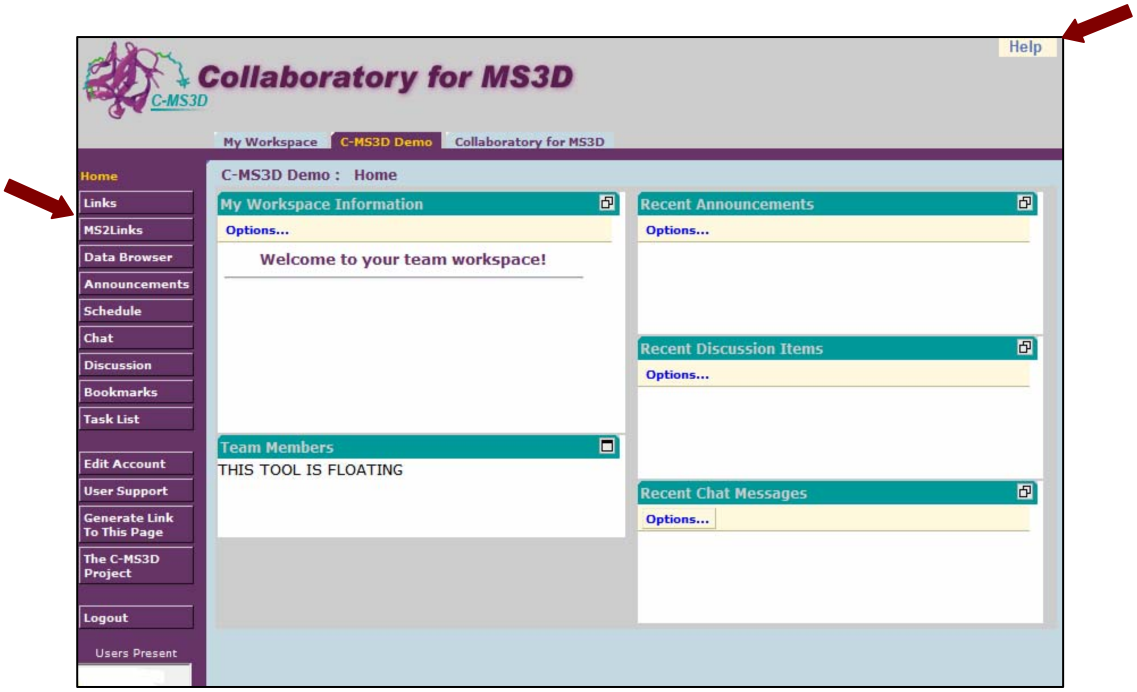
Links[9, 18] was developed at Sandia National Laboratories as a program to analyze mass spectrometric data generated from native, modified, and crosslinked protein and nucleic acid digests. MS peak lists generated from chemical crosslinking or modification experiments, followed by proteolysis and mass spectrometry. Links calculates the theoretical crosslinking and modification possibilities for single or multiple protein, RNA and/or DNA sequence(s) given information about the crosslinkers and proteases/nucleases used and the expected amino acid or base modifications. Links then returns putative assignments within a defined mass error threshold for a list of input mass (MH+) values. An analogous program, MS2Links, is used for assigning tandem MS peak lists generated from the fragmentation of crosslinked, modified or unmodified peptides, proteins and/or nucleic acids. MS2Links[18, 19] calculates the theoretical MS/MS fragment library given information about the identity of the base ion, crosslinkers (if applicable), desired ion types, and amino acid modifications. MS2Links then returns assignments within a defined mass error threshold for the list of input mass (MH+) values.

Availability

As part of the C-MS3D (Collaboratory for MS3D) initiative, web-based implementation of these programs has been made available in our development portal (https://cmcs-dev.ca.sandia.gov). To be able to access and use the programs, one needs to create an account and email us (etyu@sandia.gov) so we can add you to the MS3D team. Under the C-MS3D Demo team tab, buttons that connect to the LINKS and MS2LINKS programs will appear at the left hand side of the webpage.

* We will migrate the tools to the production portal (https://cms3d.ca.sandia.gov) in the future.

For general directions on how to navigate the portal, please refer to the portal manual. It can be accessed by clicking on the Help button at the upper right hand corner, then choose “How to”.



Welcome to your team workspace!

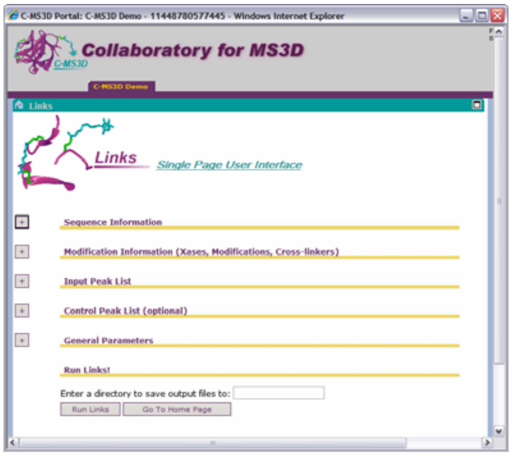
THIS TOOL IS FLOATING

II. Portal Structure

General GUI

The steps in performing a Links session are standard, however different user interfaces are made available for Links and MS2Links, affording flexibility to users.

Single Page view


C-MS3D Portal: C-MS3D Demo - 11448780577445 - Windows Internet Explorer
Collaboratory for MS3D
C-MS3D
C-MS3D Demo
Links
Links
Single Page User Interface
Run Links
Go To Home Page

Sequence Information

Modification Information (Xases, Modifications, Cross-linkers)

Input Peak List

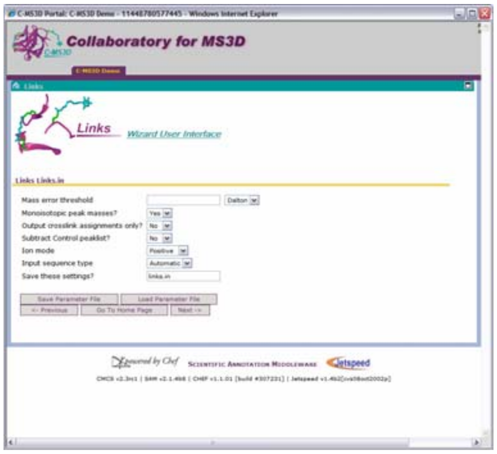
Control Peak List (optional)

General Parameters

Run Links!

Enter a directory to save output files to:

Wizard view

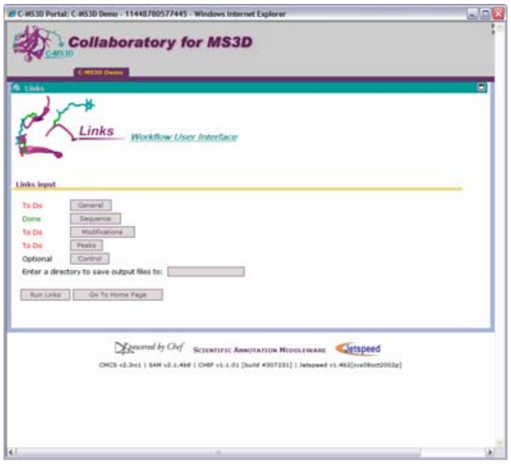


Links Links.in

Mass error threshold	Dalton
Monoisotopic peak masses?	Yes
Output crosslink assignments only?	No
Subtract Control peaklist?	No
Ion mode	Positive
Input sequence type	Automatic
Save these settings?	links.in

CMCS v2.3rc1 | SAM v2.1.4b8 | CHEF v1.1.01 [build #307231] | Jetspeed v1.4b2[rvel8oct2002p]

Workflow view



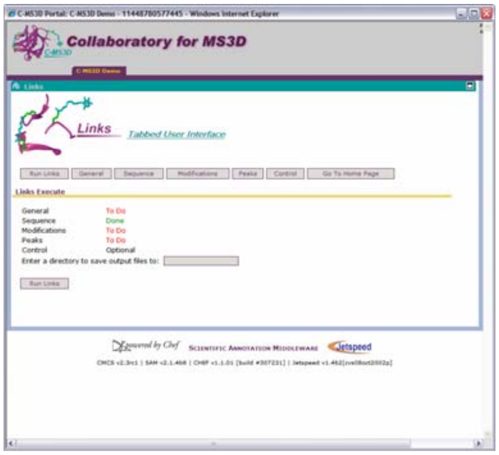
Links Input

To Do	General
Done	Sequence
To Do	Modifications
To Do	Peaks
Optional	Control

Enter a directory to save output files to:

CMCS v2.3rc1 | SAM v2.1.4b8 | CHEF v1.1.01 [build #307231] | Jetspeed v1.4b2[rvel8oct2002p]

Tabbed view



Links Execute

General	To Do
Sequence	Done
Modifications	To Do
Peaks	To Do
Control	Optional

Enter a directory to save output files to:

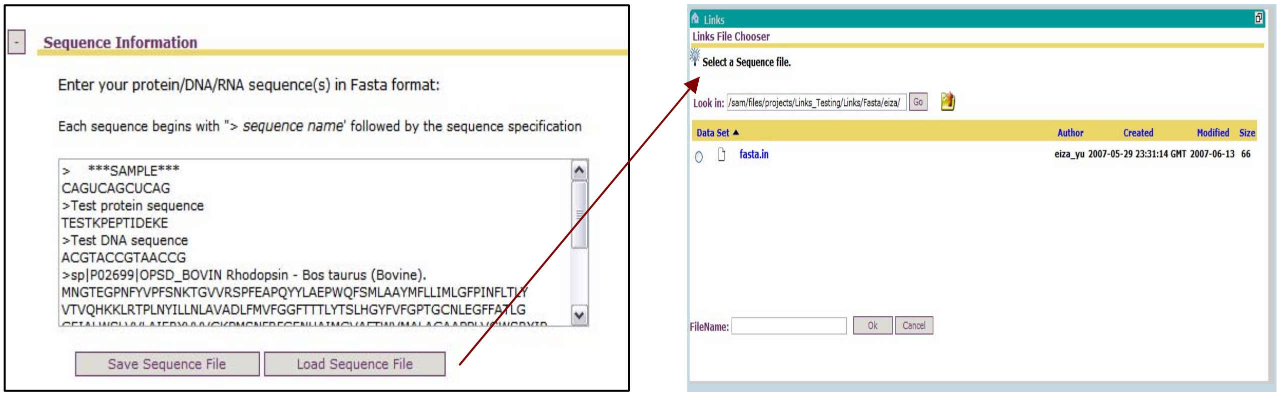
CMCS v2.3rc1 | SAM v2.1.4b8 | CHEF v1.1.01 [build #307231] | Jetspeed v1.4b2[rvel8oct2002p]

Input/Output File structures and the Data Browser

To run Links, the following input files are needed:

INPUT files	Description
Sequence.fasta	Sequence/s of protein/nucleic acids to be analyzed
Modification _Table.txt	List of modifications/ cleavages to be applied in the analysis.
Peaklist.txt	List of [M+H] observed in the mass spectra
Contro.txt	List of [M+H] observed in the control mass spectra (optional)
Links.in	Defines analysis parameters

The input files can be saved by two ways: 1) copied in the input boxes and saved through the GUIs; or 2) uploaded and saved in the Links directory through the Data Browser. For more details on the input files, please refer to section III.



Enter your protein/DNA/RNA sequence(s) in Fasta format:

Each sequence begins with "> *sequence name*' followed by the sequence specification

```
>    ***SAMPLE***
CAGUCAGCUCAG
>Test protein sequence
TESTKPEPTIDEKE
>Test DNA sequence
ACGTACCGTAACCG
>sp|P02699|OPSD_BOVIN Rhodopsin - Bos taurus (Bovine).
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
```

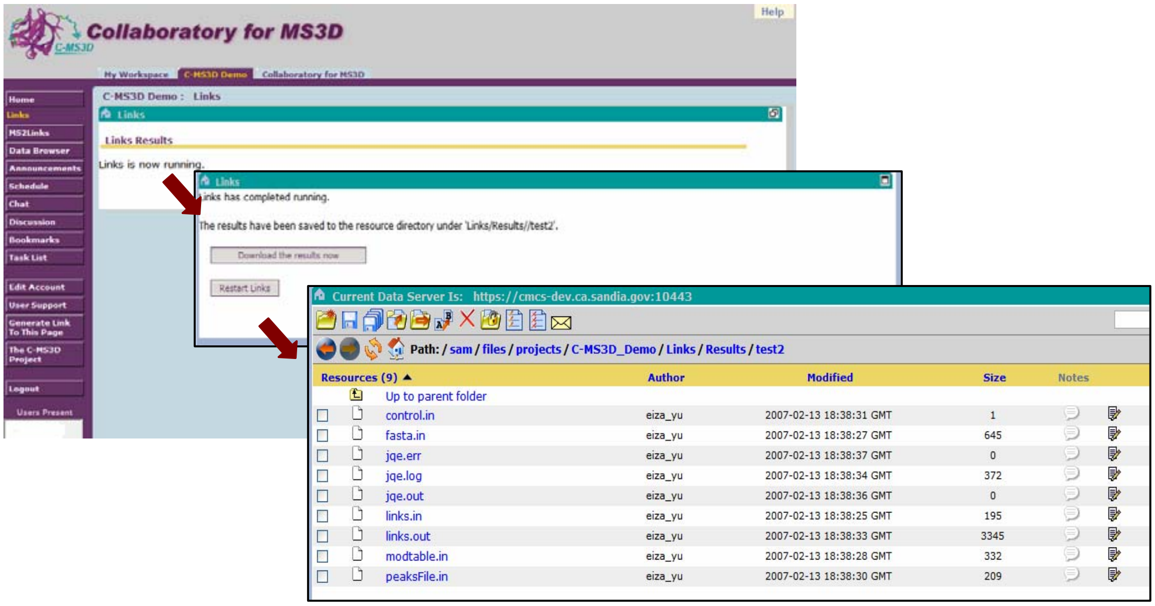
Select a Sequence file.

Look in: /sam/files/projects/Links_Testing/Links/Fasta/eiza/

Data Set ▲	Author	Created	Modified	Size
○ fasta.in	eiza_yu	2007-05-29 23:31:14 GMT	2007-06-13	66

FileName:

Once all necessary files are collected, Links will run the job through the portal, and direct the user to the Data Browser, where all the output files and associated input files are saved under one folder. Users also have the option to save their files in their own directories. If one needs to control access to their files, permissions can be set through the 🔒 icon.



Links is now running.

Links has completed running.

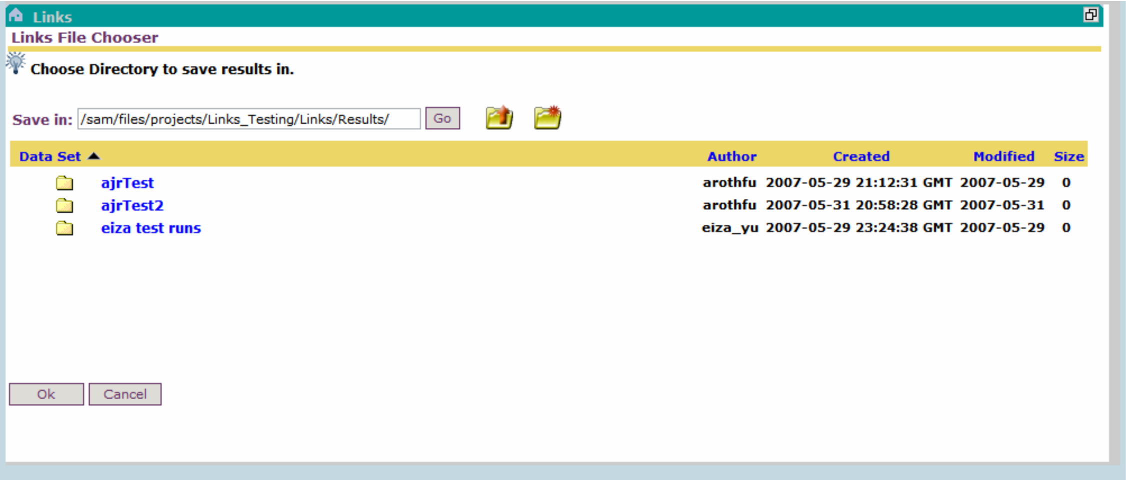
The results have been saved to the resource directory under 'Links/Results//test2'.

Current Data Server Is: https://cmcs-dev.ca.sandia.gov:10443

Path: / sam / files / projects / C-MS3D_Demo / Links / Results / test2

	Resources (9) ▲	Author	Modified	Size	Notes
	Up to parent folder				
□	control.in	eiza_yu	2007-02-13 18:38:31 GMT	1	
□	fasta.in	eiza_yu	2007-02-13 18:38:27 GMT	645	
□	jqe.err	eiza_yu	2007-02-13 18:38:37 GMT	0	
□	jqe.log	eiza_yu	2007-02-13 18:38:34 GMT	372	
□	jqe.out	eiza_yu	2007-02-13 18:38:36 GMT	0	
□	links.in	eiza_yu	2007-02-13 18:38:25 GMT	195	
□	links.out	eiza_yu	2007-02-13 18:38:33 GMT	3345	
□	modtable.in	eiza_yu	2007-02-13 18:38:28 GMT	332	
□	peaksFile.in	eiza_yu	2007-02-13 18:38:30 GMT	209	

Save results in own folder:



Choose Directory to save results in.

Save in: /sam/files/projects/Links_Testing/Links/Results/

Data Set ▲	Author	Created	Modified	Size
ajrTest	arothfu	2007-05-29 21:12:31 GMT	2007-05-29	0
ajrTest2	arothfu	2007-05-31 20:58:28 GMT	2007-05-31	0
eiza test runs	eiza_yu	2007-05-29 23:24:38 GMT	2007-05-29	0

Reporting Bugs

If you are having problems with Links/MS2Links, please refer to the FAQ section first to see if there is a quick fix to your problem. If it is not covered there or in this manual, please email us through the **User Support Form.**

Citing Links/MS2Links

Fabris D, Hawkins A, Kuntz ID, Rahn LA, Rothfuss A, Pancerella CM, Sale K, Young MM, Yang C and Yu E. The Collaboratory for MS3D: A New Cyberinfrastructure Supporting the Structural Elucidation of Biological Macromolecules and their Assemblies Using Mass Spectrometry-based Approaches.(manuscript in preparation)

Links and MS2Links were developed from the original ASAP and MS2Assign software:

Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, and Dollinger G. High throughput protein fold identificationby using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc Natl Acad Sci USA* **2000**, 97, (11), 5802-6.

Schilling B, Row RH, Gibson BW, Guo X, and Young MM. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J Am Soc Mass Spectrom* **2003**, 14, (8), 834-50.

Kellersberger KA, Yu E, Kruppa GH, Young MM, and Fabris D. Top-down characterization of nucleic acids modified by structural probes using high resolution tandem mass spectrometry and automated data interpretation. Anal Chem **2004**, 76, (9), 2438-45.

III. Running Links

1. Sequence File

The target sequence can be entered directly in the sequence editor box using the standard one-letter code (use capital letters only), with each sequence starting with a sequence name (“>name”).

```
>testprotein
TESTKPEPTIDEKE
```

The input sequence can be saved as a *.fasta file for future use. Alternatively, protein, DNA, and RNA sequences can be downloaded from databases in FASTA format, and loaded into the sequence editor box. When analyzing inter-molecular crosslinks, multiple sequences need to be entered into the sequence editor box and saved as a single fasta file. Links will automatically assign sequence numbers to the individual entries in the order that it appears in the file (See Modification Table).

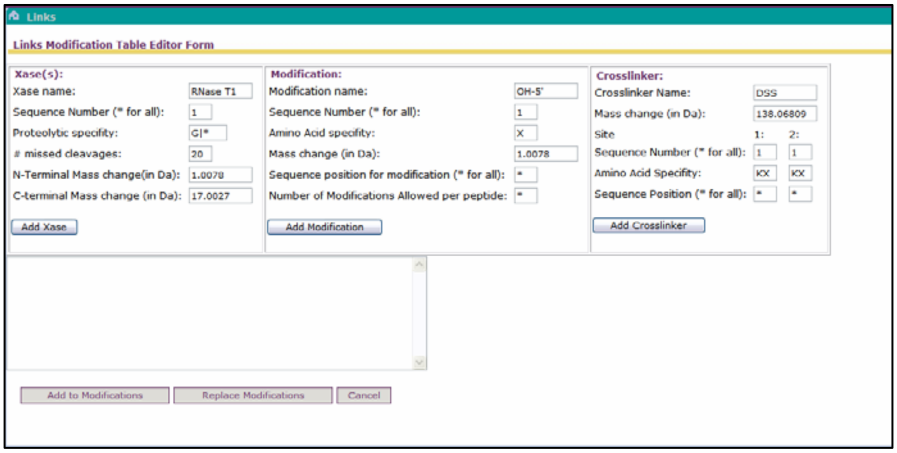
2. Modification Table

Links is capable of handling a variety of sequence modifications commonly encountered in protein and nucleic acid research. The user can define 1) amino acid-, 2) position-, or 3) peptide- specific modifications, as well as custom proteases/nucleases (XASES) and/or chemical crosslinkers, which are summarized and saved in a custom modification tables. Multiple modifications can also be defined in single session.

The Modification Table consists of the following sections:

1. TERMINAL MODIFICATIONS (applied to the termini of proteins and nucleic acids)
2. XASES (Protease,nuclease or custom cleavage applied to sequence)
3. NUCLEIC ACID MODIFICATIONS (applied to nucleic acid residues)
4. PROTEIN MODIFICATIONS (applied to amino acid residues)
5. CROSSLINKERS (generated from protein-protein or protein-nucleic acid crosslinking)

An easy way to generate the MOD Table is to use the Editor Form (click on the “Generate Modification Table” button). Modifications can be defined with the drop-down menu and added to the Mod_Table with the “Add” button.



Links Modification Table Editor Form

Xase(s):
Xase name: RNase T1
Sequence Number (* for all): 1
Proteolytic specifity: G|*
# missed cleavages: 20
N-Terminal Mass change(in Da): 1.0078
C-terminal Mass change (in Da): 17.0027

Modification:
Modification name: OH-5'
Sequence Number (* for all): 1
Amino Acid specifity: X
Mass change (in Da): 1.0078
Sequence position for modification (* for all): *
Number of Modifications Allowed per peptide: *

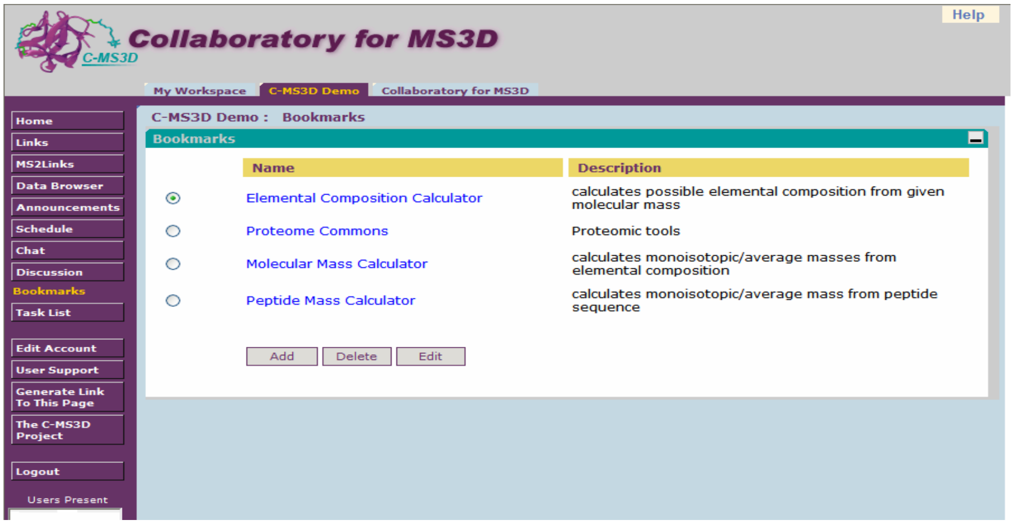
Crosslinker:
Crosslinker Name: DSS
Mass change (in Da): 138.06809

Site	1:	2:
Sequence Number (* for all):	1	1
Amino Acid Specifity:	KX	KX
Sequence Position (* for all):	*	*

Alternatively, the users can write their own modification table definitions with any text editor. Lines/entries in the modification table preceded by a # sign, indicates that these are comments or flags not in use. To define a modification, remove the # sign before the desired modification flag.

```
######### MODIFICATION TABLE##########################
XASE    *     RK|*    10       0               0       Tryp
#XASE   *     DE|*    10       0               0       GluC
MOD    *      MW     15.994914     *      *      ox-M
#MOD   *      KX     226.077589    *      3      BTm
XLINK  *      XK     *      XK     *      *      138.0681     DSS
MOD    *      KX     156.0786      *      *      DSSOH
```

If you need to add custom modifications, refer to the following sections below for more details. If a modification is not available, one can create their custom Xase/Modification/Crosslinker". To generate the needed monoisotopic and average mass shifts for the custom modification, we have provided links to peptide and molecular mass calculators through the Bookmarks tabs .



C-MS3D Demo : Bookmarks

Bookmarks

	Name	Description
●	Elemental Composition Calculator	calculates possible elemental composition from given molecular mass
○	Proteome Commons	Proteomic tools
○	Molecular Mass Calculator	calculates monoisotopic/average masses from elemental composition
○	Peptide Mass Calculator	calculates monoisotopic/average mass from peptide sequence

a. Defining terminal modifications

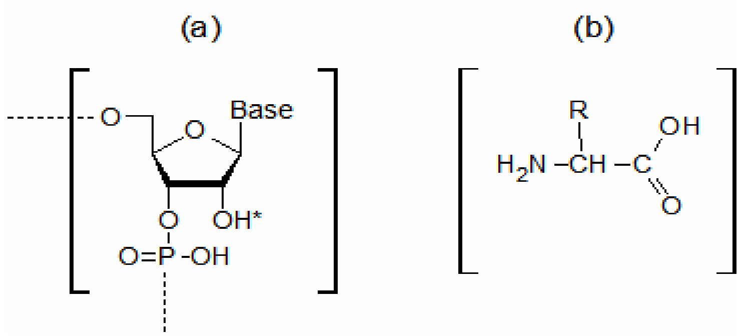
Protein and/or nucleic acid terminal modifications can be defined in END MODIFICATIONS.

END MODIFICATION Fields:
(1) Flag definition- MOD
(2) Sequence number to which MOD will be applied to.
To apply the MOD to all sequences, indicate (*) or (all).
(3) Sequence position to apply MOD
X for N-term or 5'-end
O for C-term or 3'-end
(4) Delta mass to apply (refer to basic unit figure below)
(5) Sequence position(s) to apply modification to (*=all)
(6) Number of modifications allowed per peptide
(7) A short description of the modification

```
###### END MODIFICATIONS ##################################
#MOD    1      X       1.0078          *       *       OH-5'
#MOD    1      X       80.9741         *       *       p-5'
MOD     1      X       240.9067        *       *       PPP-5'
#MOD    1      O       17.0027         *       *       p-3'
MOD     1      O       -62.9635        *       *       OH-3'
#MOD    2      X       42.0106         *       *       Acetyl-N'
```

In the example above, sequence 1 in the sequence.fasta input will have a triphosphate in the 5' end and a free hydroxyl in the 3' end. If the modifications are not provided in the list, the user can create any custom modification and add it in. To calculate delta mass (column 4), please refer to the figures below.

PROTEIN AND NUCLEIC ACID BASIC UNITS



Basic (a) nucleic acid and (b) amino acid unit. * For DNA, 2' OH is replaced by 2' H.

b. Defining sequence cleavages

Protein and/or nucleic acid digestion/cleavages can be defined in XASES.

XASE Fields:
(1) Flag definition- XASE
(2) Sequence number to which XASE will be applied to.
To apply the MOD to all sequences, indicate (*) or (all).
(3) Cleavage specificity string
(4) Number of allowed missed cleavages
(5) N-terminal/ 5'-end mass shift (after cleavage)
(6) C-terminal/ 3'-end mass shift (after cleavage)
(7) A short description of the modification

```
######    XASES##################################
#XASE   1     G|*     10     1.0078       17.0027      RNase T1
XASE    1     UC|*    10     1.0078       17.0027      RNase A
#XASE   1     *|O      5     1.0078       17.0027      5'-3'exo
#XASE   1     X|*      5     1.0078       17.0027      3'-5'exo
XASE    2     RK|*     3       0                0      Tryp
#XASE   *     FYW|*    3       0                0      Chymotryp
#XASE   *     R|*      3       0                0      Arg-C
#XASE   *     DE|*     3       0                0     Glu-C
#XASE   *     E|*      3       0                0      V8-E
#XASE   2     M|*      1       0        -48.00337      CNBr
#XASE   1     X|*      3       0                0      Carboxypeptidase
#XASE   1     *|O      3       0                0      Aminopeptidase
```

In the example above, sequence 1 in the fasta input will be digested with RNAse A with a maximum of 5 missed cleavages and sequence 2 will be cleaved with trypsin (with a maximum of 3 missed cleavages).

The syntax to handle exceptions is the " ^ " (caret) symbol. To define trypsin cleavage rules with the exception (cleaves after K,R except if followed by a P on the C-term end):

```
######    XASES##################################
XASE   2    RK|*^P    3    0         0    Tryp
```

Users can specify multiple XASES to apply to 1 sequence (i.e. Trypsin/V8 combination) in a single Links session. The user can also create any custom XASE and add it to the list. To calculate associated mass shifts, please refer to the basic unit figures above.

⚐ When performing enzymatic digestion after modification or crosslinking will result to more missed cleavages and should be accounted for in the analysis. If a lot of missed cleavages are defined, LINKS will take more time to run the job. If possible, try to limit the allowed missed cleavages < 5.

c.Defining base or amino acid residue modifications

Native and non-native modifications of proteins and nucleic acids can be defined in BASE or PROTEIN MODIFICATIONS.

MODIFICATION Fields:
(1) Flag definition- MOD
(2) Sequence number to which MOD will be applied.
To apply the MOD to all sequences, indicate (*) or (all).
(3) Amino acid or base specificity string
(4) Delta mass to apply (incremental mass shift due to modification)
(5) Sequence position(s) to apply modification to (*=all)
(6) number of modifications allowed/peptide
(7) A short description of the modification

```
######  BASE MODIFICATIONS ##################################
#MOD     1    GCA      14.0156         *       3       DMS
#MOD     *    A        73.0289         all     1       DEPC
######  PROTEIN MODIFICATIONS ##################################
#MOD     1    M        15.9949         all     5         ox-M
#MOD     1    ST       27.9949         all     5         formylation
#MOD     1    XNKSTO   14.0156         all     10        met
#MOD     1    C        24.9952         all     10        CN
MOD      2    STY      79.9663         52      1         PO4- (phosphorylation)
#MOD     1    Y        79.9568         all     10        sul- (sulfation)
#MOD     1    Y        44.9851         all     5         nit- (nitration)
#MOD     1    C        125.0477        all     10        nem-C (N-ethylmaleimide)
MOD      2    K        42.0106         all     5         ace-K (acetylation)
```

In the example above, sequence 2 will have a one phosphorylation at position 52 and will look for all possible lysine acetylations. The user can specify multiple modifications in one session.

d. Defining crosslinked substrates

Protein and/or nucleic acid crosslinking products can be defined in CROSSLINKERS.

CROSSLINKERS Fields:
(1) Flag definition- XLINK
(2) Sequence number to which XLINK will be applied.
To apply the XLINK to all sequences, indicate (*) or (all).
(3) Amino acid specificity string for site 1
(4) Sequence number
(5) Amino acid specificity string for site 2
(6) Sequence positions for site 1 (*=all)
(7) Sequence positions for site 2 (*=all)
(8) Mass change observed upon crosslinking
(9) Desciption/Name of crosslinker

```
###### CROSSLINKERS ###########################################
XLINK  *     KX     *      KX    *      *     138.06809     DSS
#XLINK *     KX     *      KX    *      *     138.06809     BS3
#XLINK *     KX     *      KX    *      *     173.9809      DTSSP
```

In the example above, Links will assign all possible DSS crosslinks between peptides generated from sequences in the fasta (In fact, Links can assign intra-molecular as well as inter-molecular crosslinked peptides, provided two sequences were entered in the fasta file and are defined in the parameter file(links.in)).

To limit the assignments to only inter-molecular crosslinks, one needs to specify the sequence numbers, as shown below as well as define it in the links parameter file (links.in):

```
###### CROSSLINKERS ###########################################
XLINK  1     KX     2      KX    *      *     138.06809     DSS
#XLINK *     KX     *      KX    *      *     138.06809     BS3
#XLINK 1     KX     2      KX    *      *     96.02113      DSG
```

In addition, the user can further limit the crosslinking assignments, if the sequence position for 1 site is known:

```
###### CROSSLINKERS ###########################################
#XLINK 2     KX     2      KX    *      *     113.99531     DST
XLINK  1     KX     2      C     *      54    165.04259     GMBS
#XLINK 2     KX     2      EDO   *      *     -18.01056     EDC
```

In this example, Links will assign GMBS crosslinks generated between C54 of protein 2 with all possible lysines in protein 1.

In practice, one can assign type 0,1, and 2 crosslinks in a single MS by having the following mod table entries:

```
######DSS CROSSLINKING MODIFICATIONS##################################
XASE    *      RK|*     3       0     0         Tryp
MOD     *     KX        156.08   *    5         DSS-mono
XLINK  1      KX         2       KX     *     *  138.06809     DSS
```

3. Input and Control Peaklist

Links needs a peaklist.txt file with the following format:

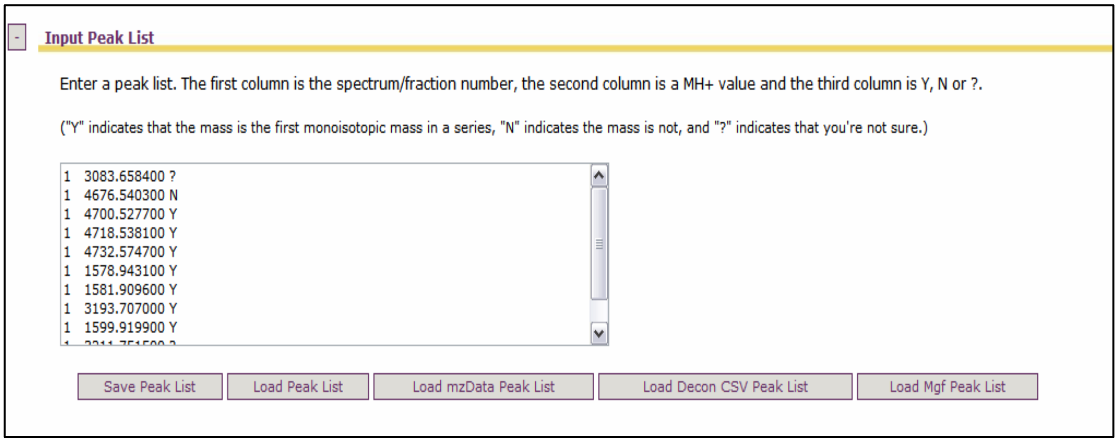
```
1      957.72       Y       1.851947e+07
1      964.01       Y       1.731260e+07
1      1000.71      Y       1.470965e+06
1      1034.74      Y       3.203310e+06
1      1056.04      Y       1.163770e+06
1      1153.88      Y       3.469220e+06
```

```
1      1169.91      Y       2.888720e+06
1      1199.87      Y       5.985100e+06
1      1213.5       Y       4.883800e+06
1      1213.5       ?       1.126710e+07
1      1225.95      N       3.124410e+06
```

The first column is the spectrum/fraction number, second column is the reduced mass (M+H or M-H), the third column indicates whether input masses are considered monoisotopic, and the last column is the intensity. For example, "Y" indicates that the monoisotopic peak was observed in the spectra. "N" indicates that the monoisotopic peak was definitely not observed (thus the mass is the C13 peak in the spectra), and "?" indicates when the monoisotopic peak was not observed due to high baseline noise, this is especially true with clusters of higher charge states at high m/z values).[18] The significance of the 3rd column values will be discussed later in the links parameters section. More columns/data can be added to the peaklist file like charge, m/z, reduced monoisotopic mass, etc, however, the program only needs the minimum peaklist information shown above.

We have created translators to extract and reformat peaklists as links input files. The users only need to upload their peaklist into the portal. We currently support the following formats:

- *.mzData
- *.mgf
- *.csv from Decon2LS (http://ncrr.pnl.gov/software/Decon2LS.stm)



Enter a peak list. The first column is the spectrum/fraction number, the second column is a MH+ value and the third column is Y, N or ?.

("Y" indicates that the mass is the first monoisotopic mass in a series, "N" indicates the mass is not, and "?" indicates that you're not sure.)

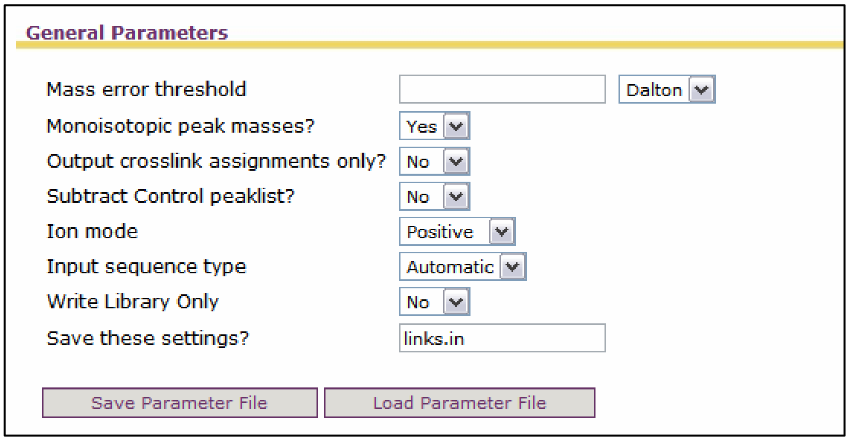
```
1   3083.658400 ?
1   4676.540300 N
1   4700.527700 Y
1   4718.538100 Y
1   4732.574700 Y
1   1578.943100 Y
1   1581.909600 Y
1   3193.707000 Y
1   1599.919900 Y
```

As a last resort, if the users peaklist information is not saved in the above formats, one can modify their M+H/M-H peaklist using Excel to add the pertinent columns and save the output as .txt files.

The control peaklist has the same format and is optional. The user can input MS peaklists obtained from control reactions, which will then be subtracted from the input peaklist and will not be considered further in the assignment stage (differential MS analysis).

4. General Parameters (Links.in)

The Links.in file defines all the parameters needed for the automated assignment of digest peaks by Links.



General Parameters

Mass error threshold		Dalton
Monoisotopic peak masses?	Yes	
Output crosslink assignments only?	No	
Subtract Control peaklist?	No	
Ion mode	Positive	
Input sequence type	Automatic	
Write Library Only	No	
Save these settings?	links.in	

a. Mass error threshold:

Links will assign all possible matches for the experimental masses provided the values are within the maximum error threshold that the user has defined. The mass error can either be in parts per million (ppm) or in Daltons (Da). The accuracy in the calibration of the mass spectrum can guide the users as to what is an acceptable and practical error threshold value to define. Having a higher mass error threshold (~100 ppm) will make the search less stringent, but is necessary when the MS was obtained from lower resolution mass spectrometers.

If the peaklist contains MH+ values with questionable monoisotopic peaks:

```
1      1213.5       ?       1.126710e+07
```

Links will try to match the experimental mass with both the theoretical monoisotopic (C12) mass as well as the C13 mass.

b. Peaklist (Monoisotopic vs average mass list):

If the peaklist contains monoisotopic masses, select yes. This depends on the resolution of the user's mass spectrometer. Otherwise, selecting "No" means that the library of theoretical masses will be calculated and matched based on average mass.

c. Output crosslink assignments only?

If the user is interested only in the assigned crosslinked species, then by selecting "Yes", Links will only write these assignments in the output.

d. Subtract Control peaklist?

Automated differential MS analysis can be performed when a control peaklist is available. Briefly, duplicate peaks will be subtracted from the input peaklist and will not be considered further in the assignment stage.

e. Ion Mode?

One can select between two modes (positive or negative) depending on which ion mode was used to obtain the mass spectra.

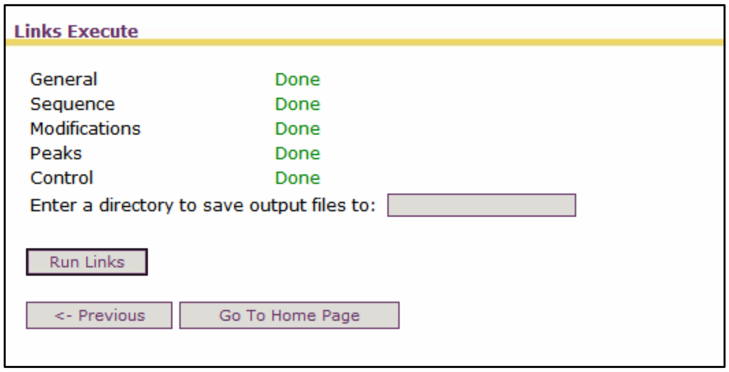
f. Input sequence type?

Links can recognize proteins, RNA, and DNA sequences automatically.

g. Write library only?

By selecting “Yes”, the user can generate the entire theoretical mass library that was defined with the given sequence file and modification table. Links writes out the theoretical mass list in the output and does not perform the matching/assignment sub-routine.

All these parameter settings will be saved in a links.in file. Once all the input files are defined, Links will prompt the user to enter a directory name for the folder where all associated files will be saved.



Links Execute

General	Done
Sequence	Done
Modifications	Done
Peaks	Done
Control	Done

Enter a directory to save output files to:

Then, one can start the Links session/job by clicking the “RUN Links” button. The output files will can be accessed from the Data Browser (please refer to the section on Output files). If the job takes longer than a couple of minutes to run, the user can log out from the portal (the job will not be terminated) and will be notified by email when the job is done.

IV. Running MS2Links

MS2Links can be used to assign mass spectrum obtained from tandem mass spectrometry of entire proteins, nucleic acids, as well as digests. Similar to Links, MS2Links also has the ability to assign putative modifications (native, chemically modified, and crosslinks).

1. Sequence file

The sequence file requirements are similar to that of Links. The sequences of an entire protein, DNA or RNA can be inputs for MS2Links. For putative crosslinked peptide species (for example); the sequences of each peptide needs to be defined.

```
>testpeptide1
TESTKPEPTIDEKE

>testpeptide2
ACRSLINKPEP
```

2. Modification Table

MS2Links modification table also follows the same requirements as that of Links. For example, a putative crosslink between peptides n the sequence fasta file can defined as:

```
######DSS CROSSLINKING MODIFICATIONS##################################
XLINK  1      KX         2       KX    *      *  138.06809    DSS
```

⚐ While the general parameters file takes care of the typical sequence ions generated from tandem MS, custom fragmentations[ref] due to a particular modification can be defined in the mod table. A well documented example, tandem MS of nucleic acids generally result to loss of nucleobases to form abasic sites and the presence of base modifications make this fragmentation more favorable.[18, 23, 24]

```
MOD     1    G     -151.049    *       1       -Gbase
```

Fragmentation of modified and/or crosslinked species do give rise to typical sequence ions, however, other fragmentation patterns seem to dominate the tandem MS. To get more details on the fragmentation behavior of such modified peptides, please refer to the following studies.[25-31]

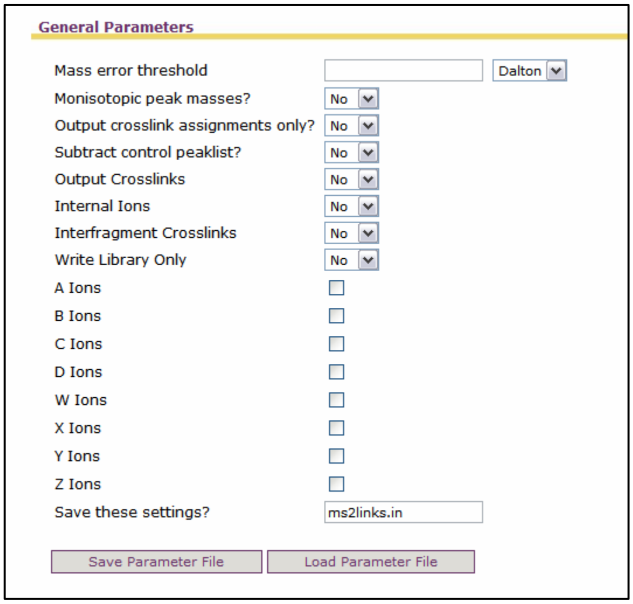
```
  MOD 1  K  140.084  *   *   DSSL13
  MOD 1  K  224.165  *   *   DSSLint
```

3. Input and Control Peaklist

Please refer to the Links input and control peaklist section as the file requirements are similar.

4. General Parameters (MS2Links.in)

The MS2Links.in file defines all the parameters needed for the automated assignment of tandem MS peaks.



General Parameters

Mass error threshold		Dalton
Monisotopic peak masses?	No	
Output crosslink assignments only?	No	
Subtract control peaklist?	No	
Output Crosslinks	No	
Internal Ions	No	
Interfragment Crosslinks	No	
Write Library Only	No	
A Ions	□	
B Ions	□	
C Ions	□	
D Ions	□	
W Ions	□	
X Ions	□	
Y Ions	□	
Z Ions	□	
Save these settings?	ms2links.in	

a. Mass error threshold:

MS2Links will assign all possible matches for the experimental masses provided the values are within the maximum error threshold that the user has defined. The mass error can either be in parts per million (ppm) or in Daltons(Da).

b. Peaklist (Monoisotopic vs average mass list):

If the peaklist contains monoisotopic masses, select yes. This depends on the resolution of the user's mass spectrometer. Otherwise, selecting "No" means that the library of theoretical masses will be calculated and will be matched based on average mass.

c. Output crosslink assignments only?

If the user is interested only in the assigned crosslinked species, then by selecting “Yes”, Links will only write these assignments in the output.

d. Subtract Control peaklist?

Automated differential MS analysis can be performed when a control peaklist is available. Briefly, duplicate peaks will be subtracted from the input peaklist and will not be considered further in the assignment stage.

c. Output Crosslinks

By selecting “Yes”, the user can generate the theoretical mass library that was defined with the sequence files and the modification tables. MS2Links writes out the theoretical mass list in the output and does not perform the matching/assignment.

d. Internal Ions

MS2Links gives the user the option to include internal sequence ion arising from tandem MS.

e. Interfragment crosslinks

By selecting “Yes”, ions generated from interfragment crosslinked species will also be considered.

f. Sequence Ion types

MS2Links gives the user the option to include all possible sequence ion types arising from tandem MS. For example, for low energy CID, the main ion types to consider are b- and y- ions. For more details, excellent reviews have been written about tandem MS of proteins[30, 32] and nucleic acids.[33, 34]

In addition to the ion types selected above, MS2Links will automatically assign water and NH_3 losses from fragment ions.

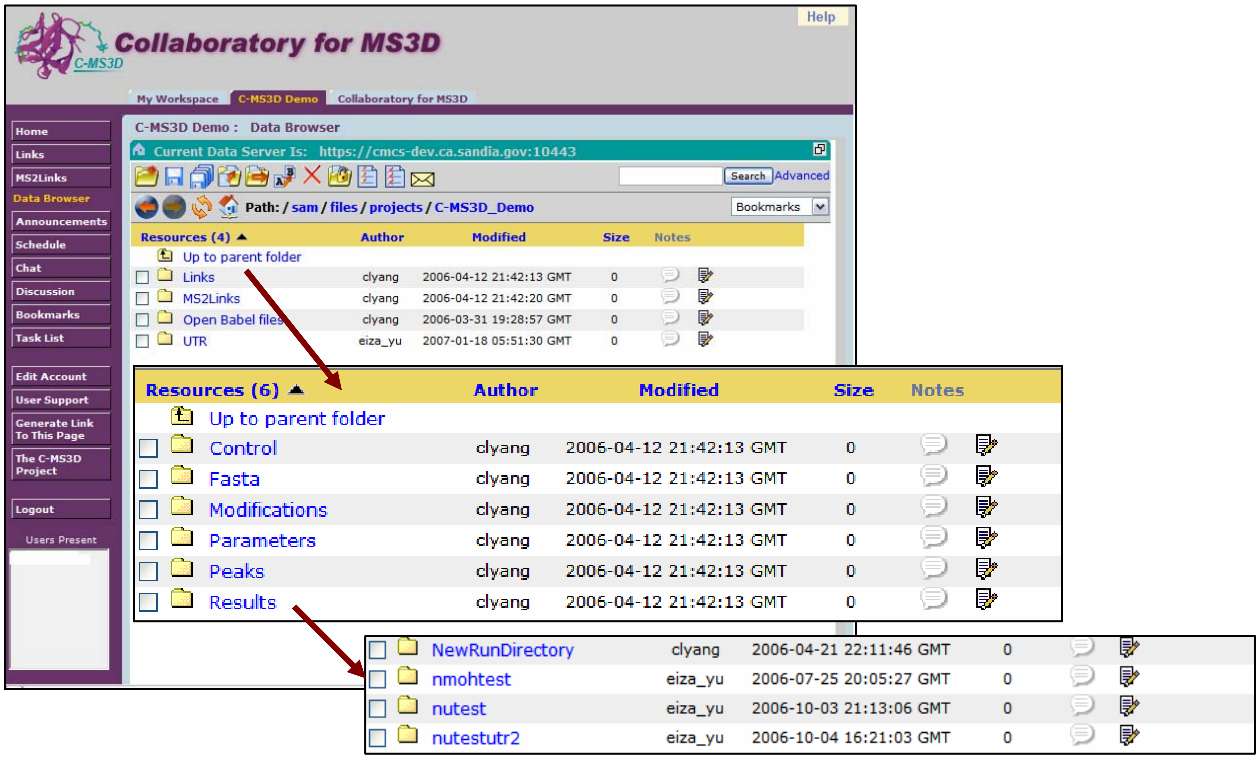
g) Write library?

By selecting “Yes”, the user can generate the entire theoretical mass library that was defined with the sequence files and the modification tables. Links writes out the theoretical mass list in the output and does not perform the matching/assignment.

All these parameter settings will be saved in a ms2links.in file. Once all the input files are defined, MS2Links will prompt the user to enter a directory name for the folder where all associated files will be saved. Then, one can start the MS2Links session/job by clicking the “RUN MS2Links” button. The output files will can be accessed from the Data Browser (please refer to the section on Output files). If the job takes longer than a couple of minutes to run, the user can log out from the portal (the job will not be terminated) and will be notified by email when the job is done.

V. Output files

Each Links/MS2Links session generates a directory in the C-MS3D portal Data Browser. At the moment, the “Download results now” button does not direct the user to the Data Browser directory.



C-MS3D Demo : Data Browser

Current Data Server Is: https://cmcs-dev.ca.sandia.gov:10443

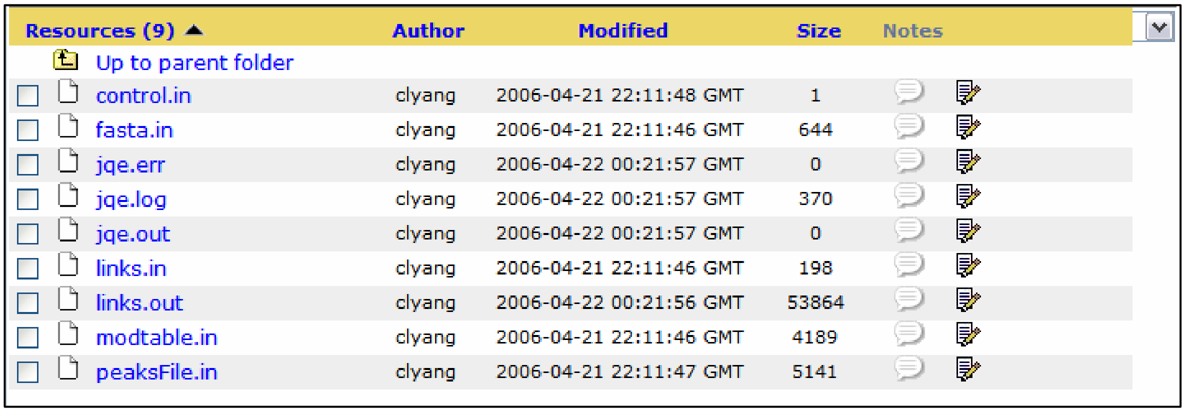
Path: / sam / files / projects / C-MS3D_Demo

	Resources (4) ▲	Author	Modified	Size	Notes
	Up to parent folder				
□	Links	clyang	2006-04-12 21:42:13 GMT	0	
□	MS2Links	clyang	2006-04-12 21:42:20 GMT	0	
□	Open Babel files	clyang	2006-03-31 19:28:57 GMT	0	
□	UTR	eiza_yu	2007-01-18 05:51:30 GMT	0	

	Resources (6) ▲	Author	Modified	Size	Notes
	Up to parent folder				
□	Control	clyang	2006-04-12 21:42:13 GMT	0	
□	Fasta	clyang	2006-04-12 21:42:13 GMT	0	
□	Modifications	clyang	2006-04-12 21:42:13 GMT	0	
□	Parameters	clyang	2006-04-12 21:42:13 GMT	0	
□	Peaks	clyang	2006-04-12 21:42:13 GMT	0	
□	Results	clyang	2006-04-12 21:42:13 GMT	0	

□	NewRunDirectory	clyang	2006-04-21 22:11:46 GMT	0	
□	nmohtest	eiza_yu	2006-07-25 20:05:27 GMT	0	
□	nutest	eiza_yu	2006-10-03 21:13:06 GMT	0	
□	nutestutr2	eiza_yu	2006-10-04 16:21:03 GMT	0	

The directory folder contains copies of all the input files used in the Links session as well as runtime log and error files (jqe.*). The jqe.* files are logs of the actual links/ms2links program execution. So if something goes wrong, they can be used to help figure out what happened.



	Resources (9) ▲	Author	Modified	Size	Notes
	Up to parent folder				
□	control.in	clyang	2006-04-21 22:11:48 GMT	1	
□	fasta.in	clyang	2006-04-21 22:11:46 GMT	644	
□	jqe.err	clyang	2006-04-22 00:21:57 GMT	0	
□	jqe.log	clyang	2006-04-22 00:21:57 GMT	370	
□	jqe.out	clyang	2006-04-22 00:21:57 GMT	0	
□	links.in	clyang	2006-04-21 22:11:46 GMT	198	
□	links.out	clyang	2006-04-22 00:21:56 GMT	53864	
□	modtable.in	clyang	2006-04-21 22:11:46 GMT	4189	
□	peaksFile.in	clyang	2006-04-21 22:11:47 GMT	5141	

The output file, links.out can be opened with a text editor or Microsoft Excel. The first part of the output file provides information on the defined parameters, sequence and modifications in the session:

```
Using MONOISOTOPIC masses.
Error is in PPM.
Reporting all assignments.
Amt of allowed error    : 50.000

LINKS: Automatic MS spectrum assignment program for cross-linked
macromolecules

Modifications:
MW        15.995       0      3     ox-M
KX       156.079       0      3     DSS-OH

Cross-link information: DSS(0-0)
            site1=XK
            site2=XK
            pos1=0
            pos2=0
            mass change = 138.068

Xase(s):
            Tryp (0)    : RK|*
            Mass mod<< 0.000
            Mass mod>> 0.000
            Missed cleavages: 5

2 modifications read in.

>APE
Length: 318
MPKRGKKGAVAEDGDELRTEPEAKKSKTAAKKNDKEAAGEGPALYEDPPD
QKTSPSAKPATLKICSWNVDGLRAWIKKKGLDWVKEEAPDILCLQETKCS
ENKLPAELQELPGLSHQYWSAPSDKEGYSGVGLLSRQCPLKVSYGIGDEE
HDQEGRVIVAEFDSFVLVTAYVPNAGRGLVRLEYRQRWDEAFRKFLKGLA
SRKPLVLCGDLNVAHEEIDLRNPKGNKKNAGFTPQERQGFGELLQAVPLA
DSFRHLYPNTPYAYTFWTYMMNARSKNVGWRLDYFLLSHSLLPALCDSKI
RSKALGSDHCPITLYLAL

Attempting to autodetect sequence type...

Input : protein sequence

Free memory4 sequences read into a 6348 member library...
```

The next part of the output file shows the assignments:

```
Expno ExpMass C13 Thr Mass Err Seq #   Seq(s) Int. Sum M.I. Charge Chrg,m/z
1 653.3089 --  0.0000 0.0   ---                                        1.851947e+07 --
1 717.3847-- 717.3934 12.2   1    80-85              GLDWVK             1.470965e+06 --
1 823.3776-- 823.3738 4.6    1    188-193            WDEAFR             1.731260e+07 --
1 839.3725—-839.3687 4.5    1    188-193 +ox           WDEAFRK             3.469220e+06 --
1 1137.5901--1137.5903 0.2    1    126-136             EGYSGVGLLSR        3.124410e+06 --
1 1403.7732-- 1403.8083 25.1  1-1   1-3,195-202        MPK-FLKGLASR       4.883800e+06 --
1 1403.7732-- 1403.8262 37.7  1-2   26-31,74-78    SKTAAK-AWIKK       5.985100e+06  --
1 2206.0691-- 2206.0259 19.6   1    86-103+1DSS-OH     EEAPDILCLQETKCSENK 3.203310e+06  --
1 2719.4036(C13)2718.3152 31.3 1-1    1-4,86-103      MPKR-EEAPDILCLQETKCSENK 5.565100e+06-
……
Number of peaks assigned = 61/119 = 51.261 percent.
```

The values for ExpNo, ExpMass, C13, Int.Sum, M.I., Charge, chrg, m/z columns are all copied from the input peaklist.in file(if available).

The Links output file:

If the experimental mass is monoisotopic, the C13 column will show no value. However, in the case below,

```
1 2719.4036(C13)2718.3152 31.3 1-1    1-4,86-103      MPKR-EEAPDILCLQETKCSENK 5.565100e+06
```

the experimental mass was matched to the (C13) value of the theoretical mass.

The sequence identifiers Seq# and Seq(s) provide the identity of the protein in the fasta file as well as the amino acid/nucleotide sequence of the assignment. For example;

```
1 717.3847-- 717.3934 12.2   1    80-85              GLDWVK             1.470965e+06 --
```

shows that the assigned peptide mass is from position 80-85 with sequence GLDWVK of sequence 1 in the fasta.in file.

Crosslinked species are reported as (position#proteinA, position#proteinB) in the Seq(s) column. In the example below, the assignment is for an **intra-molecular** crosslink in protein 1, with position 1-3 crosslinked to 195-202.

```
1 1403.7732-- 1403.8083 25.1  1-1   1-3,195-202        MPK-FLKGLASR       4.883800e+06 –
```

The example below shows an assignment for an **inter-molecular** crosslink between position 26-31 of protein 1, and position 74-78 of protein 2.

```
1 1403.7732-- 1403.8262 37.7  1-2   26-31,74-78    SKTAAK-AWIKK       5.985100e+06  --
```

Modified species are reported as (position# + MOD) in the Seq(s) column. In the example below, the assignment is for an oxidized W residue in protein 1.

```
1    188-193 +ox            WDEAFRK
```

This is also the same format for reporting intra-peptide crosslinks:

```
1-1            26-32 + DSS            SKTAKWK
```

The MS2Links output file:

The MS2Links output follows the same conventions as Links regarding crosslinks and modifications. As for the crosslinked sequence ions, MS2Links follows the same convention set in MS2Assign.[19]

VI. References

1. Maier, C. S.; Deinzer, M. L., Protein conformations, interactions, and H/D exchange. *Methods Enzymol* **2005,** 402, 312-60.

2. Bahar, I.; Erman, B.; Haliloglu, T.; Jernigan, R. L., Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry* **1997,** 36, (44), 13512-23.

3. Steinhoff, H. J., Inter- and intra-molecular distances determined by EPR spectroscopy and site-directed spin labeling reveal protein-protein and protein-oligonucleotide interaction. *Biol Chem* **2004,** 385, (10), 913-20.

4. Bettio, A.; Beck-Sickinger, A. G., Biophysical methods to study ligand-receptor interactions of neuropeptide Y. *Biopolymers* **2001,** 60, (6), 420-37.

5. Lorenz, M.; Diekmann, S., Distance determination in protein-DNA complexes using fluorescence resonance energy transfer. *Methods Mol Biol* **2006,** 335, 243-55.

6. Back, J. W.; de Jong, L.; Muijsers, A. O.; de Koster, C. G., Chemical cross-linking and mass spectrometry for protein structural modeling. *J Mol Biol* **2003,** 331, (2), 303-13.

7. Rappsilber, J.; Siniossoglou, S.; Hurt, E. C.; Mann, M., A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal Chem* **2000,** 72, (2), 267-75.

8. Sinz, A., Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom Rev* **2006,** 25, (4), 663-82.

9. Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; Dollinger, G., High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc Natl Acad Sci U S A* **2000,** 97, (11), 5802-6.

10. Yu, E.; Fabris, D., Direct probing of RNA structures and RNA-protein interactions in the HIV-1 packaging signal by chemical modification and electrospray ionization fourier transform mass spectrometry. *J Mol Biol* **2003,** 330, (2), 211-23.

11. Yu, E. T.; Zhang, Q.; Fabris, D., Untying the FIV frameshifting pseudoknot structure by MS3D. *J Mol Biol* **2005,** 345, (1), 69-80.

12. Robinette, D.; Neamati, N.; Tomer, K. B.; Borchers, C. H., Photoaffinity labeling combined with mass spectrometric approaches as a tool for structural proteomics. *Expert Rev Proteomics* **2006,** 3, (4), 399-408.

13. Bennett, K. L.; Matthiesen, T.; Roepstorff, P., Probing protein surface topology by chemical surface labeling, crosslinking, and mass spectrometry. *Methods Mol Biol* **2000,** 146, 113-31.

14. Urlaub, H.; Hartmuth, K.; Luhrmann, R., A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods* **2002,** 26, (2), 170-81.

15. Guan, J. Q.; Chance, M. R., Structural proteomics of macromolecular assemblies using oxidative footprinting and mass spectrometry. *Trends Biochem Sci* **2005,** 30, (10), 583-92.

16. Steen, H.; Jensen, O. N., Analysis of protein-nucleic acid interactions by photochemical cross-linking and mass spectrometry. *Mass Spectrom Rev* **2002,** 21, (3), 163-82.

17. de Koning, L. J.; Kasper, P. T.; Back, J. W.; Nessen, M. A.; Vanrobaeys, F.; Van Beeumen, J.; Gherardi, E.; de Koster, C. G.; de Jong, L., Computer-assisted mass spectrometric analysis of naturally occurring and artificially introduced cross-links in proteins and protein complexes. *Febs J* **2006,** 273, (2), 281-91.

18. Kellersberger, K. A.; Yu, E.; Kruppa, G. H.; Young, M. M.; Fabris, D., Top-down characterization of nucleic acids modified by structural probes using high-resolution tandem mass spectrometry and automated data interpretation. *Anal Chem* **2004,** 76, (9), 2438-45.

19. Schilling, B.; Row, R. H.; Gibson, B. W.; Guo, X.; Young, M. M., MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J Am Soc Mass Spectrom* **2003,** 14, (8), 834-50.

20. Peri, S.; Steen, H.; Pandey, A., GPMAW--a software tool for analyzing proteins and peptides. *Trends Biochem Sci* **2001,** 26, (11), 687-9.

21. Seebacher, J.; Mallick, P.; Zhang, N.; Eddes, J. S.; Aebersold, R.; Gelb, M. H., Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. *J Proteome Res* **2006,** 5, (9), 2270-82.

22. Tang, Y.; Chen, Y.; Lichti, C. F.; Hall, R. A.; Raney, K. D.; Jennings, S. F., CLPM: a cross-linked peptide mapping algorithm for mass spectrometric analysis. *BMC Bioinformatics* **2005,** 6 Suppl 2, S9.

23. Andersen, T. E.; Kirpekar, F.; Haselmann, K. F., RNA fragmentation in MALDI mass spectrometry studied by H/D-exchange: mechanisms of general applicability to nucleic acids. *J Am Soc Mass Spectrom* **2006,** 17, (10), 1353-68.

24. Kirpekar, F.; Krogh, T. N., RNA fragmentation studied in a matrix-assisted laser desorption/ionisation tandem quadrupole/orthogonal time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **2001,** 15, (1), 8-14.

25. Gaucher, S. P.; Hadi, M. Z.; Young, M. M., Influence of crosslinker identity and position on gas-phase dissociation of Lys-Lys crosslinked peptides. *J Am Soc Mass Spectrom* **2006,** 17, (3), 395-405.

26. Fenaille, F.; Tabet, J. C.; Guy, P. A., Study of peptides containing modified lysine residues by tandem mass spectrometry: precursor ion scanning of hexanal-modified peptides. *Rapid Commun Mass Spectrom* **2004,** 18, (1), 67-76.

27. Back, J. W.; Hartog, A. F.; Dekker, H. L.; Muijsers, A. O.; de Koning, L. J.; de Jong, L., A new crosslinker for mass spectrometric analysis of the quaternary structure of protein complexes. *J Am Soc Mass Spectrom* **2001,** 12, (2), 222-7.

28. Bakhtiar, R.; Guan, Z., Electron capture dissociation mass spectrometry in characterization of peptides and proteins. *Biotechnol Lett* **2006,** 28, (14), 1047-59.

29. Larsen, M. R.; Trelle, M. B.; Thingholm, T. E.; Jensen, O. N., Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *Biotechniques* **2006,** 40, (6), 790-8.

30. Wells, J. M.; McLuckey, S. A., Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **2005,** 402, 148-85.

31. Gorman, J. J.; Wallis, T. P.; Pitt, J. J., Protein disulfide bond determination by mass spectrometry. *Mass Spectrom Rev* **2002,** 21, (3), 183-216.

32. Paizs, B.; Suhai, S., Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **2005,** 24, (4), 508-48.

33. Crain, P. F., Mass spectrometric techniques in nucleic acid research. *Mass Spectrom Rev* **1990,** 11, (1), 3-40.

34. Nordhoff, E.; Kirpekar, F.; Roepstorff, P., Mass spectrometry of nucleic acids. *Mass Spectrom Rev* **1996,** 15, (2), 67-138.